aurum

AURUM MÜHENDİSLİK SİSTEMLERİ VE MİMARLIK DERGİSİ
AURUM JOURNAL OF ENGINEERING SYSTEMS AND ARCHITECTURE

Cilt 2, Sayı 1 | Bahar 2018
Volume 2, No. 1 | Spring 2018, 27-31

WEB BASED PROGRAM FOR BIG MOLECULAR DATA CONVERSION FOR ANALYSIS BY MATLAB, PHYTON OR R

Halit Hami OZ[1]
[1]Kafkas University, Faculty of Engineering and Architecture, Department of Computer Engineering, Kars, 36100, Turkey
hamioz@yahoo.com

Mahmut AYDIN[2]
[2]Kafkas University, Faculty of Engineering and Architecture, Department of Computer Engineering, Kars, 36100, Turkey
mahmut083@hotmail.com

Abstract

Molecular data is created in different formats: MICROSAT, SNP, AFLP, RFLP, DNA, RNA, DISTANCE, PROTEIN, DART, INDEL, RAPID. File formats includes Arlequin, Genpop, Structure, Nexus, Mega, Fasta. Scientists working in this field needs to analyze this molecular data, and he/she does it by either writing special programs to convert these big data to the format he needs or sends his/her data to some centers for analysis. User friendly and easy to use web based molecular data converting program was deveopled using R programming language at Kafkas University Department of Bioengineering and Department of Computer Engineering. Users can upload their data using the Web based program selecting input and out file formats to convert their big molecular data to the format they want for analysis using either R, Phyton programming languages or MATLAB Wavelet Toolbox™

Keywords: Matlab, Wavelet, Big Data, R, Bioenformatics.

MATLAB, PHYTON VEYA R KULLANARAK WEB TABANLI BÜYÜK MOLEKÜLER VERİ DÖNÜŞÜM ANALİZİ

Özet

Moleküler veriler farklı formatlarda oluşturulur: MICROSAT, SNP, AFLP, RFLP, DNA, RNA, MESAFANT, PROTEİN, DART, INDEL, HIZLI. Dosya biçimleri arasında Arlequin, Genpop, Structure, Nexus, Mega, Fasta bulunur. Bu alandaki bilim adamları bu moleküler veriyi analiz etmek ve bu büyük veriyi ihtiyaç duyduğu biçime dönüştürmek için özel programlar yazarak ya da verilerini analiz için bazı merkezlere göndererek yapar. Kullanıcı dostu ve kullanımı kolay web tabanlı moleküler veri dönüştürme programı, Kafkas Üniversitesi Biyomühendislik ve Bilgisayar Mühendisliği Bölümü'nde R programlama dili kullanılarak geliştirilmiştir. Kullanıcılar, giriş ve çıkış dosya formatlarını seçen Web tabanlı programı kullanarak veriler yükleyebilirler ve R programlama dili veya MATLAB Wavelet Toolbox kullanarak verilerini analiz edebilirler.

Anahtar Kelimeler: Matlab, Wavelet, Big Data, R, Bioenformatics.

1. INTRODUCTION

Junior scientists who are working in Molecular Biology and Genetics and Bioengineering at the Faculty of Engineering and Faculty of Medicine wanted to learn Python and R programming languages to analyse their big data. They said when they need to cooperate with the international scientists on projects they all have been asked whether they knew Python and R programming languages. When they wanted to attend some workshops to learn to develop applications to analyze big data, prior knowledge of Python was always a prerequisite to attend R programming languages workshops. They said they need to learn these programming languages to analyze their big biodata which sometimes as much as terabyte in size. Some of them attended my Matlab programming course 3 hours/week for one semester to learn Matlab before attending R programming language course since there was no Python course available. R language is currently used by most of the molecular biology scientists around the world to analyze their big data.

Molecular data is created in different formats: MICROSAT, SNP, AFLP, RFLP, DNA, RNA, DISTANCE, PROTEIN, DART, INDEL, RAPID. File formats includes Arlequin, Genpop, Structure, Nexus, Mega, Fasta. Scientists working in this field needs to analyze this molecular data, and he/she does it by either writing special programs to convert these big data to the format he needs or sends his/her data to some centers for analysis. There are also some programs available to convert the same molecular data, but none of them is web based. Therefore, we decided to develop one web based user friendly program for converting big biodata.

Materials and Method

A web based molecular data converter program using R programming language was developed at Kafkas University Department of Bioengineering and Department of Computer Engineering. Users can upload their data using the Web based program selecting input and out file formats to convert their big biodata to the format they want.

Results

The developed web based big biodata converter program is very user friendly and easy to use. There are also other programs available to convert the same molecular data, but none of them is web based. Now the program goes through field testing and the work is still in progress.

Discussion

There are several programs/programming languages used to analyze wavelet data. Scientests working in wavelet are using different applications and one of the most common applications is Wavelet Toolbox of MatlabTM program. The other programming languages are Fortran, IDL (Interactive Data Language),and Python. Torrence and Combo (1998), Artail, et al. (2014), Bruce and Gao (1996) have some guidelines for wavelet analysis.

Scientist who wants to analyze wavelet data needs to use one of these programming languages and she/he also needs to know how to write codes in these programming languages. However, the very same

scientist who is working with wavelet may not be very good at writing the codes. It takes years of hard work to learn one programming language and write the code to solve a problem. Some sample wavelet analysis programs written in Matlab, Fortran, IDL and Python are give in the following web site Wavelet Software: http://paos.colorado.edu/research/wavelets/software.html

Matlab has "Wavelet Toolbox™ provides functions and apps for analyzing and synthesizing signals, images, and data that exhibit regular behavior punctuated with abrupt changes. The toolbox includes algorithms for the continuous wavelet transform (CWT), scalograms, and wavelet coherence. It also provides algorithms and visualizations for discrete wavelet analysis, including decimated, nondecimated, dual-tree, and wavelet packet transforms. In addition, you can extend the toolbox algorithms with custom wavelets". http://www.mathworks.com/products/wavelet/?requestedDomain=www.mathworks.com#)

"The toolbox lets you analyze how the frequency content of signals changes over time and reveals time-varying patterns common in multiple signals. You can perform multiresolution analysis to extract fine-scale or large-scale features, identify discontinuities, and detect change points or events that are not visible in the raw data. You can also use Wavelet Toolbox to efficiently compress data while maintaining perceptual quality and to denoise signals and images while retaining features that are often smoothed out by other techniques" (http://www.mathworks.com/products/wavelet/?requestedDomain=www.mathworks.com#)

MATLAB is used as a neural network tool (Poojitha et al. 2016), analysis of bio-signals using wavelet transform and genetic algorithm (Sukiennik and Bialasiewicz, 2015). "The Multivariate Exploratory Data Analysis (MEDA) Toolbox in Matlab is a set of multivariate analysis tools for the exploration of data sets. Multivariate Statistical Process Control (MSPC) charts and data simulation/approximation algorithms (ADICOV) are also included in the toolbox. Most of the exploratory tools are extended for their use with very large data sets (Big Data), with unlimited number of observations" (Camacho et al. 2015).

"PCA toolbox for MATLAB is a collection of modules for calculating Principal Component Analysis, as well as Cluster Analysis and Multidimensional Scaling, which are two other wellknown multivariate methods for unsupervised data exploration" (Ballabio, 2015). Rodrigues et al. (2015) claim that their own implementation "DataIP outperforms MatLab and R by several orders of magnitude when it comes to handling large number of instances or large number of parameters".

Lai (2015) constructed a "simple power system model in DIgSILENT PowerFactory which performs the short circuit analysis for the High Impedance Fault(HIF) and the open circuit islanding for the heavy load conditions". He used MATLAB for discrete wavelet transform analysis for the output from the power system model.

Rübel et al. (2013) developed "OpenMSI, a software framework and platform for efficient access, management, and analysis of the data generated by Mass spectrometry imaging (MSI). The OpenMSI file format supports storage of raw MSI data, metadata, and derived analyses in a single, self-describing format based on HDF5 and is supported by a large range of analysis software (e.g., Matlab and R) and programming languages (e.g., C++, Fortran, and Python)".

Dinc and Baleanu (2003) used MATLAB 6.5 software for one-dimensional wavelet analysis. Bigdely-Shamlo et al. (2015) propose a standardized early-stage EEG processing pipeline (PREP) and discuss the application of the pipeline to more than 600 EEG datasets . Users can download the PREP pipeline as a freely available MATLAB library from http://eegstudy.org/prepcode. Manojbhai et al. (2016) developed template using MATLAB for "high performance computing combined with analytics helps to overcome the challenges posed by Big Medical Image data. Dynamic pattern template is then used for searching relevant images from the existing repositories based on the image query."

2. CONCLUSION

Users can upload their data using the user friendly web based program selecting input and out file formats to convert their big biodata to the format they want. This converted data can later be analysed by Matlab or R and programs written in Phyton or Fortran programming languages. The developed web based program currently goes through field testing and the coversion of big biodata results obtained using this program and other programes are compared to determine the validity and reliability of the program.

3. REFERENCES

Artail, H.A., Al-Asadi, H., Koleilat, W. and Chehab, A. 2004. "Applications of a Spreadsheet-based Wavelet Analysis Toolbox in Education". *Int. J. Engeng. Ed.* 20(6), 920-927.

Ballabio, D. 2015. "A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure." *Chemometrics and Intelligent Laboratory Systems* 149,1–9.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, KM., and Robbins, K.A. 2015. "The PREP pipeline: standardized preprocessing for large-scale EEG analysis". *Frontiers in Neuroinformatics,* 9: 16.

Bruce, A. and Gao, H.Y. 2016. " Applied Wavelet Analysis with S-Plus

Springer-Verlag New York, Inc. Secaucus, NJ, USA ©1996, ISBN:0387947140 http://dl.acm.org/citation.cfm?id=547924 accessed on July 15, 2016

Camacho, J., Perez-Villegas, A., Rodriguez-Gomez, R.A., and Jiménez-Mañas, E. 2015. "Multivariate Exploratory Data Analysis (MEDA) Toolbox for Matlab". *Chemometrics and Intelligent Laboratory Systems,* 143,49-57.

Dinc¸ E. and Baleanu, D. 2003. "Multidetermination of thiamine HCl and pyridoxine HCl in their mixture using continuous daubechies and biorthogonal wavelet analysis." *Talanta,* 59(4), 707-717.

Lai, C.S. 2015. " High Impedance Fault and Heavy Load under Big Data Context", *IEEE International Conference on Systems Man and Cybernetics Conference Proceedings,* Pages: 653-658

Manojbhai, D. D., Pradipkumar, K. K., and Rajamenakshi, R. 2016. "Big Image Analysis for Identifying Tumor Pattern Similarities". *Proceedings of 2016 International Conference on Advanced Communication Control and Computing Technologies(ICACCCT),* Pages: 39-43.

Poojitha, V; Bhadauria, M., B., Shilpi, J., and Anchal Garg, A. 2016. "A collocation of IRIS Flower using Neural network Clustering tool in MATLAB." *6th International Conference on Cloud System and Big Data Engineering,* Pages: 53-58.

Rodrigues, A., Silva, C., Borges, P., Silva, S., and Dutra, I. 2015. "Performance Evaluation of Statistical Functions", *IEEE International Conference on Smart CityY/Socialcom/Sustaincom (Smartcity,)* Pages: 754-760.

Rübel, O., Greiner, A., Cholia, S., Louie, K., Bethel, E.W., Northen, T.R. and Bowen, B.P. 2013. " OpenMSI: A High-Performance Web-Based Platform for Mass Spectrometry Imaging", *Analytical Chemistry,* 85(21), 10354-10361.

Sukiennik, P and Bialasiewicz, J.T. 2015. "Cross-correlation of bio-signals using continuous wavelet transform and genetic algorithm". *Journal of Neuroscience Methods*, 247, 13-22.

Torrence, C and and Combo, G.P. 1998. "A Practical Guide to Wavelet Analysis. Bulletin of the American Meteorological Society" DOI: http://dx.doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2 Published Online: 1 January 1998 http://journals.ametsoc.org/doi/pdf/10.1175/1520-0477%281998%29079%3C0061%3AAPGTWA%3E2.0.CO%3B2 accessed on July 15, 2016

Matlab Wavelet Toolbox Documentation. http://www.mathworks.com/help/wavelet/index.html;jsessionid=da1c0f14d9f97badb7f882d5934b